ICSA

INSTITUTE FOR COMPUTER SERVICES AND APPLICATIONS

RICE UNIVERSITY

Phase III of the Rice University Earth Resources Data Analysis Program

FINAL REPORT
(June 1974 - May 1975)

Institute for Computer Services and Applications
Rice University
Houston, TX 77001
June, 1975

This research was sponsored by NASA under Contract NAS 9-12776

FINAL REPORT

I. Introduction:

During the preceding contract year, a variety of subtasks have been performed mostly in two areas: 1) systems analysis and 2) algorithmic development. The major effort in the systems analysis task (see Section II) was the development of a recommended approach to the monitoring of resource utilization data for the Large Area Crop Inventory Experiment (LACIE). Other efforts included participation in various studies concerning the LACIE Project Plan, the utility of the GE Image 100, and the specifications for a special purpose processor to be used in the LACIE. In the second task (see Section III), the major effort was the development of improved algorithms for estimating proportions of unclassified remotely sensed data. Also, work was performed on optimal feature extraction and optimal feature extraction for proportion estimation.

This report summarizes the findings of these tasks. Details of some of these tasks are to be found in ICSA technical reports referenced herein.

II. Task 1: Systems Analysis

This study developed a rationale and a method for a system to collect resource utilization (RU) data for the LACIE. The method employed for conducting this study was to adopt a "top-down" approach toward the design of such a system. The first step was to determine who would be the likely users of such data and what were the anticipated uses. Next, the types and amounts of data needed were determined from a detailed inspection of the proposed system. The last step was to devise a scheme for obtaining the data, getting it into the system, and providing for generation of appropriate reports as needed. Details of this study may be found in "The Resource Utilization Monitoring System for the Large Area Crop Inventory Experiment—a Recommend Approach" by R. A. Schafer, D. L. Van Rooy, and M. S. Lynn, ICSA Report #275-025-020 and "Data Base Design & Maintenance for the Resource Utilization Monitoring System for the Large Area Crop Inventory Experiment—a Recommended Approach," by R. A. Schafer, ICSA Report #275-025-021.

It was found that in general, the users of RU data can be classified into two groups differentially by their objectives:

1. To operationally monitor the resource utilization of the system in order to spot processing bottlenecks, improve data flow, provide audit and security facilities, etc.

2. To post-hoc examine the resource utilization of the system in order to determine the cost-effectiveness of the system and to provide appropriate data to aid in the development of similar future systems.

Any or all of the above suggested uses of RU data may not presently, or in the future, be the intention of any group now connected with the LACIE and are only suggestions as to the applicability of RU data in the two categories of usage. The presently known users of RU data in category 1 are the LACIE subsystem managers, the Earth Observations Division Management Team and the Project Management Team. The only presently known user in category 2 is the USDA, although the USDA will also be doing some operational monitoring. Other users not presently known would likely be organizations interested in performing LACIE-type functions on their own computer systems.

Five types of RU data were identified: computer usage, manpower usage, materials usage, overhead and throughput rate. Due to operational difficulties in determining a means of quantifying overhead, that type of RU data was not included in the eventual system design. The amount of data to be collected, due to the scope of the LACIE, was considerable; thus a method of organizing the raw data into a useable form was examined. A project accounting structure was decided upon as a basis for this organization. A hierarchy of reporting/accounting levels was established and an information retrieval system was described which utilized that structure.

A structured data base was designed with the lowest level being the LACIE subsystem (or easily separable functions within a subsystem) and higher levels being the geographic structure of LACIE, i.e. stratum, zone, region, and country. This data base

was designed to be kept on magnetic tape and the update procedure involved copying an old master tape onto a new master tape accumulating data from an update tape. This update tape would be created from RU data supplied to a Resource Data Manager. A set of forms for this data was described as examples of the data to be collected. A sample set of reports was also designed on the same basis. A major design objective of the data base and reporting system was to provide the flexibility to produce reports on arbitrary combinations of the RU data, since the current understanding of future needs was incomplete.

Some programs were then written to gather some of the data (in particular from the ERIPS DELOG tape). Also an information retrieval system available at Rice was used to produce sample reports using simulated data.

Rice University personnel also participated in several other systems analysis studies. These included an examination of the General Electric Image 100 for use by the EOD as an applicati[illegible] development systems; participation in the design review of the LACIE; and a critique of the specifications for the special purpose processor to be used in the LACIE.

III. Task II: Algorithmic Development

Most of the effort on this task was devoted to the development and testing of two algorithms for estimating proportions of classes contained in multispectral data. The motivation for this work comes mostly from the LACIE, where the total acreage of a crop, rather than its exact location, is one of the major quantities of interest.

The first method (see "Optimal Design of an Unsupervised Adaptive Classifier with Unknown Priors," by D. Kazakos, ICSA Report #275-025-013) involves classification of the data while updating the estimate of the proportions. To test this algorithm, a version for the special case of two classes was programmed and pseudo-random data was generated. The algorithm for this case is:

$$P_{n+1} = P_n - \frac{1}{n+1} L * (G - W)$$

where P_n is the n^{th} estimate of the prior probability of class number one,

$$L = (A - B)^{-1}$$

$$G = P_n \cdot A + (1 - P_n) * B$$

with $$A = \int_S f_1(x)\, dx$$

$$S = \left\{ x \;\middle|\; \frac{f_1(x)}{f_2(x)} \geq \frac{1 - P_n}{P_n} \right\}$$

$$\beta = \int_S f_2(x)\ dx$$

$f_i(x)$ being the density function for the i^{th} class

and

$$W = \begin{cases} 1 & \text{if } \dfrac{f_1(x_{n+1})}{f_2(x_{n+1})} \geq \dfrac{1 - P_n}{P_n} \\ 0 & \text{otherwise} \end{cases}$$

The asymptotic variance of P_n about the true proportion π is given by

$$E(P_n - \pi)^2 = \frac{1}{n} \frac{G(\pi)\ \left(1 - G(\pi)\right)}{L(\pi)^2}$$

The P_n's are then bounded in the interval $(0,1)$.

This algorithm was tested on some 2-class, one-dimensional pseudo-random data. It was found that the data needs to be taken in scrambled order, and that the algorithm has a tendency to "stick" to a boundary ($P_n \simeq 1$ or 0) in its present form. By bounding $L(P_n)$, this "sticking" could be obviated, but the variance of the estimates was still considerably larger in some cases than the asymptotic error variance, even for as many as 1000 points. These experiments led us to the conclusion that the finite sample bias of this algorithm is probably too large for use in LACIE, so further development of this algorithm was ceased.

Subsequently, a new method for proportion estimation was developed (see "Recursive Estimation of Prior Probabilities Using the Mixture Approach," by D. Kazakos, ICSA Report #275-025-019).

This algorithm uses a recursive estimate of the prior probabilities to achieve results comparable to those of maximum likelihood estimation (the results should be the same for the special case of 2 classes) though the former is much easier to implement and computationally more efficient. This algorithm was tested using both $M \geq 2$ class and $N \geq 1$ dimensional pseudo-random and Hill County LANDSAT data (the same data used in "An Empirical Comparison of Five Proportion Estimators," by W. A. Coberly and P. L. Odell, Annual Report for NASA contract NAS 9-13512 for the University of Texas at Dallas). The function $L(P_n)$ which controls the error variance was found to be too complex to evaluate at each step. Therefore, the algorithm was modified whereby a constant L was used and this constant depended only on the statistics of the classes. This approach will produce some degradation in the estimate of the proportions. Other modifications of the algorithm include the bounding and renormalizing of the current estimate of the prior probabilities at each step, and the scrambling of the order of the data so as to prevent blocks of data belonging to single classes from "confusing" the estimator.

The results of the testing on Hill County data is given in Table 1. For this case, the calculated value of L was ~ 11.5. Other values of L were also used since it was felt that L probably should be restricted to $L \gtrsim 10$. The maximum likelihood estimate obtained by Coberly and Odell for this case is also given. There is some doubt about the true proportions of wheat and barley since one "wheat" field is consistently classified as barley; so the numbers in parentheses refer to the proportions if that field really

is barley. However, the mean-squared error (MSE) is given about the other proportions. These results indicate that the recursive estimator can achieve results comparable to the maximum likelihood one, but the problem of what value to choose for L remains. Presently, we are investigating other approximations to the function L which could alleviate this difficulty. Further development and testing including a timing comparison with the maximum likelihood estimator and a further study into the requirements for shuffling the data is now underway, and a report will be issued on the results obtained. We believe that this estimator is the most promising of the two developed here and therefore recommend that all the development effort be put into this estimator rather than the first one.

A related project was some preliminary development of an algorithm for optimal feature extraction for estimating proportions. For the special case of two Gaussian classes, an expression was derived for an upper bound of the error variance when optimally estimating proportions. This bound is expressed in terms of the Bhattacharrya distance, and it was shown that maximizing the Bhattacharrya distance minimizes this bound. Thus, existing feature extraction algorithms (the University of Houston one) may be used for this special case. A report on this work is in preparation. Due to the importance of this effort for LACIE and other projects, we recommend that the EOD support or perform further research and development in this area for the more general case of m classes. This could reduce computation costs and provide bounds for the error to be used in estimating the total error in the acreage estimate.

	Wheat	Fallow	Barley	Grass	Stubble	Total MSE
	(.302)		(.179)			
True Proportions	.366	.286	.115	.079	.147	—
Maximum Likelihood	.300	.297	.177	.086	.140	.010
Recursive Estimate						
L=11.5	.286	.230	.142	.101	.241	.022
L=7	.308	.270	.164	.067	.190	.012
L=3	.313	.267	.182	.061	.177	.010
L=1	.305	.292	.192	.084	.126	.012

Hill County Data

Table I

Further development on another feature extraction algorithm took place and some initial testing was done during this contract year. This algorithm minimizes the increased risk of misclassification (see "Optimal Linear and Nonlinear Risk of Misclassification," by R. J. P. de Figueiredo, ICSA Report #275-025-014. This work has been jointly sponsored with the U. S. Army under contract DA-31-124-ARO-D-462, the U. S. Air Force under contract AFOSR-75-2777, and the NSF under grant GK-36375). Progress has been slower than expected in the testing phase, but presently the algorithm yields satisfactory results for the special case of a linear transformation from n dimensions to one. Testing will continue on another program that treats the more general $n \rightarrow k$ dimensionality reduction. A report on the results of the first program is being prepared and will be available shortly. Another report will be issued following development and testing of the second program.